Received: 14 Jul 2025 | Revised: 31 Jul 2025 | Accepted: 01 Dec 2025 | Published: 30 Dec 2025
Volume 3(2) Pages 93-109 DOI: https://doi.org/ 10.56279/jicts.v3i2.415

Journal of ICT Systems

Leveraging Tabular Variational Autoencoders for Student Loan Risk Modelling in Data-Scarce Environments

Reuben Moyo [1], Jimmy Mbelwa, Juma Lungo

Department of Computer Science and Engineering, University of Dar es Salaam, Dar es Salaam, Tanzania

[1]Reuben Moyo
Email: reubencmoyo@gmail.com

Funding information

No funding information available.

Keywords

Data Scarcity
Multi-Task Learning
Student Loan
Synthetic Data
Tabular Variational Autoencoder

Abstract

Higher Education Loan Boards (HELBs) in Low- and Middle-Income Countries (LMICs) must assess student loan repayment risk despite data scarcity, missingness, and class imbalance. Traditional models such as Logistic Regression and Random Forest often perform poorly under these conditions, limiting early risk identification. This study develops a two-stage predictive framework integrating a Tabular Variational Autoencoder (TVAE) for high-quality synthetic data augmentation and a Multi-Input Multi-Output neural architecture with Multi-Task Learning (MIMO-MTL) to jointly estimate Probability of Default (PD) and Loss Given Default (LGD). TVAE outperforms CTGAN and Gaussian Copula and was selected for its ability to preserve joint feature structure. Using data from Malawi's Higher Education Students Loans and Grants Board (HESLGB), the framework achieved an AUC of 0.96 for PD and an R^2 of 0.74 for LGD, improving risk detection and enabling earlier identification of high-risk borrowers. SHAP analysis enhanced interpretability by identifying key socioeconomic and academic rivers, offering a scalable, policy-relevant approach to strengthening student loan sustainability in a data scarce environment.

1. Introduction

Access to higher education financing in Low- and Middle-Income Countries (LMICs) is primarily facilitated through student loan programs managed by national Higher Education Loan Boards (HELBs). These agencies, such as Malawi's Higher Education Students Loans and Grants Board (HESLGB), Tanzania's Higher Education Students Loans Board (HESLB), and Kenya's Higher Education Loans Board (HELB), are legally required to promote equitable access to higher education, regardless of socioeconomic background [1, 2] . Their policies prioritise social

protection and inclusivity over credit-based assessments, which compels them to assist even high-risk or financially disadvantaged applicants [3, 4] . As a result, these loan schemes function in inherently high-risk and data-scarce environments, where incomplete financial records, volatile unemployment, and informal sector incomes are prevalent [5-7] . The sustainability of these schemes is significantly challenged by high delinquency rates, ranging from 28–35% in Malawi to 21–30% in Tanzania [8], combined with extensive missing data and extreme class imbalance. These alarming rates undermine the predictive reliability of traditional credit scoring models, making the accurate estimation of Probability of Default (PD) and Loss Given Default (LGD) a persistent challenge.

Traditional statistical and machine learning models, such as Logistic Regression, Decision Trees, Random Forest, and Gradient Boosting, face fundamental limitations in LMIC contexts. The models degrade under extreme imbalance and missingness, particularly when PD and LGD are modelled independently. Logistic Regression assumes linearity and complete feature information, which leads to poor performance in the presence of high missingness and non-Gaussian socioeconomic distributions [9]. While ensemble models like Random Forest and XGBoost enhance the handling of non-linearity, they still tend to overfit minority classes and suffer significantly under extreme imbalance [10]. Furthermore, conventional imputation techniques often fail to accurately reconstruct the complex socioeconomic interactions that drive default behaviour in datasets related to education financing. These structural limitations render traditional approaches inadequate for the operational needs of HELBs, particularly when datasets are sparse, noisy, and exhibit heterogeneous covariate relationships.

To address these challenges, recent advancements in synthetic data generation offer a promising shift in approach. Frameworks such as Synthetic Data Vault (SDV) [11], Gaussian Copula Modelling [12], Conditional Tabular GANs (CTGAN) [13], and Tabular Variational Autoencoders (TVAE) [14] are capable of producing high-fidelity synthetic samples that maintain the statistical properties of real datasets while alleviating privacy and sparsity concerns. Generative models are particularly beneficial for educational loan datasets in LMICs because they can reconstruct missing socioeconomic patterns through latent representations, oversample rare delinquency cases without altering feature distributions, stabilise predictive models trained on small or imbalanced datasets, and facilitate privacy-preserving data augmentation required by public institutions [15].

Among the various generative methods, TVAE offers several advantages. Unlike Gaussian Copula models, which assume simple Gaussian marginals, TVAE captures non-linear, high-dimensional joint distributions [16], a critical aspect for modelling interdependent academic, demographic, and socioeconomic variables. Additionally, in contrast to CTGAN, which often suffers from mode collapse when handling rare categorical combinations typical in HELB datasets, TVAE maintains distributional diversity through its latent-space learning architecture and KL-regularisation mechanisms. Its suitability for mixed-type tabular data has been demonstrated in several comparative studies [14, 17] .

Despite a growing interest in generative modelling, its application to student loan risk analysis in LMICs remains limited. Existing studies focus primarily on default classification and rarely examine LGD, despite its importance for provisioning and liquidity planning. Moreover, no

prior work integrates TVAE augmentation with Multi-Input Multi-Output (MIMO) architectures and Multi-Task Learning (MTL) to jointly estimate PD and LGD, despite these outcomes being empirically correlated and shaped by shared socioeconomic determinants. Joint PD-LGD modelling is theoretically aligned with credit risk decomposition theory [18], which recognises that PD and LGD are interdependent components of expected credit loss, rather than independent metrics, while MTL is supported by representation sharing theory in Machine Learning [19, 20] . In the context of student loan portfolios, common latent factors such as employment outcomes, macroeconomic conditions, and loan design features simultaneously affect both default likelihood and loss severity.

By jointly estimating PD and LGD through a MIMO-MTL framework, the model can effectively capture these shared risk factors while accounting for the distinct statistical properties of each task. This methodology mitigates estimation variance in situations of data scarcity and aligns predictive modelling with the underlying economic framework of unsecured credit risk. This study proposes a TVAE-augmented MIMO-MTL framework to address data scarcity, missingness, and imbalance in LMIC loan portfolios. We evaluate TVAE against SDV, CTGAN, and Gaussian Copula and develop a joint PD–LGD model whose performance and interpretability are assessed using SHAP analysis. The study contributes both (i) a generalisable methodological framework that integrates TVAE-based augmentation with joint PD–LGD modelling, and (ii) a concrete MIMO-MTL predictive model as one implementation of this framework.

2. Method

A. Dataset

This study utilised three primary data sources: (i) administrative borrower-level data obtained from HESLGB, (ii) macroeconomic indicators sourced from the National Statistics Office (NSO), and (iii) a graduate labour market survey designed to capture supplementary socioeconomic information. The core dataset was sourced from HESLGB, which maintains comprehensive administrative records related to loan disbursements, beneficiary characteristics, academic progress, and repayment histories. It comprised 350,981 records from 1992 to 2017, including both labelled and unlabelled borrower records, annual macroeconomic data aligned to repayment entry years, and survey-derived socioeconomic variables spanning four domains: demographic, socioeconomic, academic, and loan-specific attributes (Table 1).

Table 1. Structure and feature composition of the dataset.

Category	Features	Description	Missing (%)
Demographic	Age, Gender, Marital status, Region	Borrower characteristic	10-60
Socioeconomic	Unemployment-Rate, Parental Education	Economic context	30-80
Academic	College Year of Entry, Graduation Status, Major	Academic performance	10-30
Loan-specific	Loan Amount, Interest Rate, Repayment Period, Loan Type	Loan details	5-50

The dataset exhibited characteristics typical of LMIC educational loan systems: high missingness,

exceeding 60%, in income- or employment-related attributes (Table 1); severe class imbalance due to a few status records on delinquent borrowers (Figure 1); mixed data types, including continuous, ordinal, and categorical; and data entry noise due to multiple sources. Ethical approval was obtained from the Mzuzu University Research Ethics Committee (Protocol MZ/202307/03), and all processing complied with national data protection requirements.

B. Data Preprocessing

Data preprocessing is critical to ensure quality of the HESLGB dataset for modelling. The preprocessing followed the pipeline described in Figure 2. Missing numerical values were imputed using k-Nearest Neighbours (k=5), selected for its robustness to sparse data. Categorical features, such as gender and region, were imputed using mode imputation to preserve their distribution. Numerical features were normalised using MinMaxScaler to scale values between 0 and 1, ensuring compatibility with neural network training. Categorical features were One-Hot encoded to convert them into a format suitable for modelling. For class imbalance, the original delinquency rate of less than 1% severely limited model learning. Rather than applying SMOTE, which distorts high-dimensional distributions, the study adopted generative oversampling using TVAE.

C. Synthetic Data Generation Using TVAE

Given the scarcity of delinquency cases in the dataset, synthetic data were generated using a TVAE. It was selected following a comparative evaluation with SDV, Gaussian Copula, and CTGAN. It demonstrated superior preservation of joint feature dependencies, evidenced by the lowest Maximum Mean Discrepancy (MMD), and exhibited greater robustness in modelling rare categorical combinations, an area where CTGAN commonly suffers from mode collapse [17]. Unlike the Gaussian Copula, TVAE does not rely on Gaussian marginal assumptions, allowing it to generate diverse and realistic samples that better

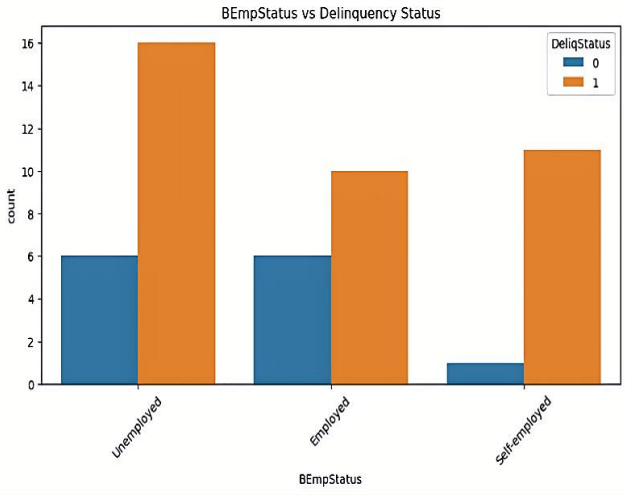


Figure 1. Delinquency rate by employment status.

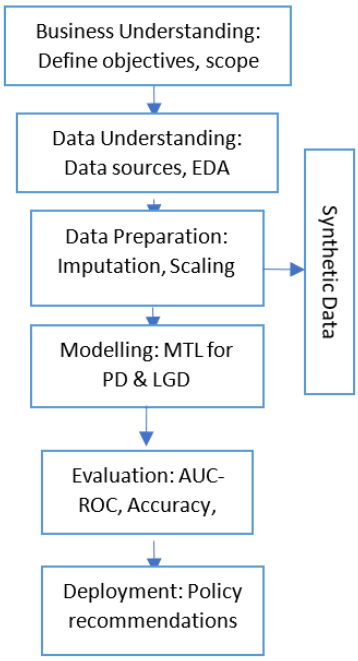


Figure 2. Research pipeline.

reflect the heterogeneity inherent in mixed-type loan data [14]. Its effectiveness for high-dimensional, heterogeneous tabular datasets has been documented in recent studies [14,16], further supporting its adoption in this work.

The TVAE framework comprises an encoder network that maps each input vector x into a latent distribution $q(z \mid x)$ and a decoder network that reconstructs synthetic observations via $p(x \mid z)$. Model training involves maximising the Evidence Lower Bound (ELBO), defined as

$$\mathcal{L}_{ELBO} = \mathbb{E}_{q(z|x)}[\log p(x \mid z)] - D_{KL}(q(z \mid x) \parallel p(z)),$$

which balances reconstruction accuracy with regularisation of the latent space. This approach enables TVAE to generate high-fidelity synthetic instances that enhance minority class representation without distorting the underlying data structure. A total of 150,000 synthetic samples were generated, yielding a final training dataset of 500,981 records. This proportion was identified during sensitivity analysis as the optimal balance between mitigating imbalance and avoiding generative noise. Critically, synthetic data was used only in the training set; all model evaluation was performed on purely real, unseen data, ensuring external validity.

D. Proposed Framework

The predictive component employs a MIMO neural network trained under an MTL framework to estimate the PD and LGD jointly. Joint modelling is theoretically justified because PD and LGD depend on overlapping socioeconomic and academic determinants when two outcomes share a latent structure (Figure 3).

MTL provides more efficient estimators and improved generalisation compared to training independent models, particularly under data sparsity.The architecture consists of a shared feature extractor with three dense layers (128, 64, and 32 units, ReLU activation), regularised with ℓ_2 weight decay (0.01) and dropout (0.3). This shared representation captures nonlinear relationships that influence both PD and LGD.

Two task-specific heads follow. The PD head uses a 16-unit dense layer with sigmoid activation and binary cross-entropy loss:

$$\mathcal{L}_{PD} = -\sum_i [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] .$$

The LGD head uses a 16-unit dense layer with linear activation and Mean Absolute Error loss:

$$\mathcal{L}_{LGD} = \frac{1}{n} \sum_i |y_i - \hat{y}_i| .$$

The optimisation objective combines the two tasks:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{PD} + \lambda_2 \mathcal{L}_{LGD},$$

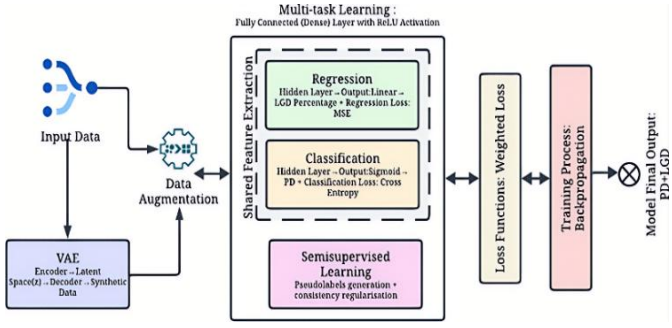


Figure 3. MIMO-MTL Framework.

with $\lambda_1 = 0.6$ and $\lambda_2 = 0.4$, selected via grid search to balance the relative task contributions. This formulation allows the model to exploit shared information between PD and LGD, demonstrating the downstream predictive advantages of training on TVAE-augmented data.

E. Model Training and Evaluation Procedure

The modelling pipeline followed established standards for credit risk evaluation to ensure methodological rigour. The dataset was partitioned into an 80% training set and a 20% test set, with synthetic data exclusively included in the training set to avoid information leakage. Stratified sampling was employed to maintain the empirical delinquency ratio, which is a best practice when working with highly imbalanced borrower-level data.

Models were trained using the Adam optimiser (learning rate = 0.001) and a batch size of 64 for up to 100 epochs. To stabilise optimisation and reduce overfitting, early stopping and adaptive learning-rate reduction techniques were implemented, reflecting standard deep learning practices for tabular risk prediction. Performance was assessed using metrics grounded in credit-scoring literature. For PD, AUC quantified discriminatory power, while F1-score evaluated classification reliability under imbalance. For LGD, R^2 and MAE measured explanatory accuracy and prediction error, consistent.

To contextualise the contribution of the proposed framework, three benchmark models were implemented: Logistic Regression, Random Forest, and XGBoost. Logistic Regression remains the regulatory and theoretical baseline in credit scoring due to its interpretability and well-understood statistical properties [21]. Random Forest represents a widely adopted non-linear ensemble approach known for its robustness to missing data [22]. XGBoost offers a robust gradient-boosting benchmark, consistently achieving state-of-the-art performance in tabular financial prediction tasks. Their inclusion ensures a theoretically grounded and empirically meaningful comparison against established modelling paradigms [23].

F. Interpretability via SHAP

To ensure transparency and policy alignment, model explanations were generated using SHAP (SHapley Additive exPlanations). SHAP quantifies each feature's contribution to PD and LGD predictions, enabling HELBs to understand underlying risk drivers, validate model behaviour, and implement fair and accountable decision-making. This is essential in public-sector credit systems where model-based decisions affect vulnerable borrowers.

G. Computational Consideration

The experiments were conducted using Python 3.8, TensorFlow 2.19, and scikit-learn 1.6.0 on a CUDA-GPU NVIDIA GeForce RTX 3050 Ti (16 GB) laptop. To ensure robustness, the model was trained five times with different random seeds, and the performance metrics were reported as means with standard deviations. Table 2 presents hyperparameters, including loss weights (λ_1,λ_2), learning rate and layer sizes, which were tuned via gridsearch on the validation set, optimised for balanced PD and LGD performance.

Table 2. Hyperparameter search space.

Hyperparameter	Search Space
Learning rate	[0.001, 0.01]
Units in classification branch	[32, 64, 128]
Units in regression branch	[16, 32, 64]
Dropout rate	0.3
L2 regularisation	[0.001, 0.01]
Batch size	[16,32]
Lambda_1, Lambda_2	[0.3-0.7,0.3-0.7]

3. Results

A. TVAE Synthetic Data Quality

Generating synthetic data similar to realistic data is a crucial task in data augmentation and data production. Despite its advantages, synthetic data generation introduces potential limitations, including the risk of distributional shifts, where generated data may fail to capture rare but critical patterns present in real delinquency cases [12]. Furthermore, generative adversarial models, such as CTGAN, may suffer from mode collapse, resulting in reduced diversity in synthetic samples [13].

To mitigate these challenges, rigorous data fidelity assessments were conducted using statistical similarity metrics to ensure that the generated datasets accurately reflect real data distributions and relationships, while maintaining the quality of the synthetic data [24]. The study utilised the Kolmogorov-Smirnov test, Pairwise Correlation Coefficient Comparison, data quality report, and diagnostic score.

The Kolmogorov–Smirnov (KS) test is a widely utilised method for assessing differences in data distributions. This test measures the distance between the empirical distribution functions of two distinct datasets [25]. Being nonparametric, it does not rely on any assumptions about the underlying distributions of the data. The KS static is expressed as

$$D = sup_x|F_1(x) - F_2(x)|,$$

Where F_1 and F_2 are the cumulative distribution functions of the real dataset and the synthetic dataset, respectively (Table 3).

Most features generated by TVAE exhibited high fidelity, with no significant distributional divergence from the real data, as confirmed by p-values from the KS test exceeding 0.05 for all features except a mild deviation of LossDefault, consistent with known challenges in modelling extreme LGD tails. The Kolmogorov-Smirnov test was used to compare the distributions of numerical features between the real and synthetic datasets (Table 3). It measured the maximum difference between the empirical distribution functions of the two datasets. On the one hand, a lower KS statistic and a high p-value indicate that the synthetic data follows a similar distribution to the real data. For instance, features like LAmount and LossDefault had KS statistics of 0.9632 and 0.2042, respectively, indicating moderate to significant distributional differences. On the other hand, features like BGradeYear and BGPA showed a low KS statistic, suggesting a closer match between the real and synthetic distributions.

Table 3. Kolmogorov-Smirnov (KS) Test Results for Synthetic Data Fidelity.

Feature	KS Statistic	p-value	JS Divergence
BAge	0.1143	0.4953	0.015
HouseholdSize	0.1773	0.0759	0.018
Dependents	0.3296	0.427	0.019
CollegeYrOfEntry	0.1035	0.6209	0.017
BGradYear	0.0566	0.9944	0.012
BGPA	0.0858	0.8243	0.048
RegionUnemplyRate	0.98	0.252	0.013
InflationRate	0.4532	0.693	0.017
GradRate	0.073	0.9352	0.016
CurrentLIntRate	0.0711	0.9465	0.038
OgLIntRate	0.0898	0.7821	0.018
LTerm	0.091	0.7685	0.014
LoanGracePeriod	0.021	0.924	0.021
LAmount	0.7632	0.833	0.019
RAmnt	0.3354	0.614	0.011
DeliqStatus	0.0669	0.9677	0.016
LossDefault	0.2042	0.0264	0.013

Although RegionUnemplyRate and LAmount exhibit higher KS statistics, these marginal differences do not materially affect downstream model performance. RegionUnemplyRate is an aggregate macroeconomic indicator characterised by regime-based clustering, while LAmount displays a highly skewed, policy-constrained distribution. TVAE moderates extreme values to

stabilise latent learning while preserving correlation structure and the joint PD–LGD relationships central to expected credit loss estimation. Low MMD supports this.

B. Correlation Structure Preservation

The study also employed a pairwise correlation coefficient comparison, an effective method for evaluating synthetic data, as it helps determine how well the relationships between features in the synthetic dataset reflects those in the original data [17, 24]. Correlation coefficients measure the strength and direction of the linear relationships between pairs of variables, providing a clear view of how features interact. Preserving these relationships is critical for ensuring that the synthetic data captures the underlying structure and patterns of the real dataset. By comparing the correlation matrices of the real and synthetic datasets, it becomes possible to identify how closely the synthetic data replicates these inter-variable connections. Significant differences in correlation values suggest that the synthetic data may not accurately reflect the key patterns of the original data, which could limit its usefulness for tasks such as predictive modelling or statistical analysis [14].

Various metrics were employed to assess data quality, including Jensen-Shannon Divergence, which measures the similarity between distributions with lower values indicating a better fit. The Diagnostic Score evaluates the accuracy of a logistic regression classifier trained on synthetic data and tested on real data. Additionally, MMD quantifies divergence in high-dimensional feature space using a Gaussian kernel. Lastly, the Quality Score is calculated as a weighted average of KS statistics (0.4), correlation preservation (0.4), and feature overlap (0.2).

To assess the relationships between features, pairwise correlation coefficients were calculated for both the real and synthetic datasets. The goal was to verify that the synthetic data preserved the interdependencies between variables, which are crucial for predictive modelling. The correlation comparison revealed that, although the overall trend was preserved, certain variables exhibited more substantial differences between real and synthetic correlations. These discrepancies may impact downstream models, necessitating additional fine-tuning of the synthetic data generation process. Table 4 presents a comparison of the first four features and the last four features.

Table 4. Pairwise correlation comparison between real and synthetic data.

Feature	Real Data	Synthetic Data	Difference
BAge	1.00	1.00	0.00
Household	-0.02	0.01	0.03
Dependents	-0.22	-0.34	0.12
RTime	0.01	0.12	0.11
LAmount	-0.22	-0.01	0.21
RAmount	-0.92	-0.91	0.01
DeliqStatus	0.08	0.11	0.03
lossDefault	1.00	1.00	0.00

C. Comparative performance of Generative Models

Synthetic data generation was employed using three distinct generative models: TVAE, CTGAN, and Gaussian Copula. The rationale for experimenting with multiple methods was to comparatively assess their ability to capture and reproduce the complex statistical properties inherent in student loan delinquency and LGD data. Triangulating the approaches allowed for a robust evaluation of each model's effectiveness in preserving data fidelity, managing class imbalance, and enhancing model performance, ultimately informing the selection of the most suitable technique for improving the accuracy and reliability of predictive models developed in this study.

To evaluate the synthetic data quality, multiple metrics were employed (Table 4), including Mean Kolmogorov-Smirnov (KS) Statistic, Mean

Absolute Error (MAE), Quality Score, and Diagnostic Score. The Mean KS Statistic measures the maximum difference between the cumulative distribution functions of real and synthetic data for numerical features, with lower values indicating closer alignment of the distributions. MAE quantifies the average absolute difference between real and synthetic data for continuous features. The Quality Score is a composite metric that evaluates the overall statistical similarity between real and synthetic datasets, based on a weighted combination of KS statistics, correlation preservation, and feature distribution overlap. The Diagnostic Score assesses the utility of synthetic data for downstream predictive tasks by measuring the performance of a simple classifier trained on synthetic data and tested on real data, with higher scores indicating better utility.

Table 5. Summary of evaluation of the model synthesisers.

Model	Mean KS Statistic	MAE	Quality Score	Diagnostic Score	Data Fidelity (MMD)
Gaussian Copula	0.21	0.31	0.83	0.98	0.018
CTGAN	0.23	0.32	0.80	0.98	0.025
TVAE	0.25	0.31	0.77	0.98	0.012

To further contextualise TVAE's performance, it was compared against CTGAN and Gaussian Copula using additional metrics tailored to tabular data (Table 5). Metrics are interpreted hierarchically: KS and JS assess marginal plausibility, while MMD evaluates joint distribution fidelity, which is theoretically aligned with multi-output risk modelling. MMD measures the divergence between real and synthetic data distributions in a high-dimensional feature space, with lower values indicating higher fidelity. Tabular Suitability is a qualitative assessment based on the model's ability to handle mixed data types (discrete and continuous) and preserve complex, non-Gaussian distributions, rated as High, Moderate, or Low based on expert evaluation and performance on benchmark tabular datasets [14]. While TVAE's Quality Score (0.77) is slightly lower than Gaussian Copula (0.83) and CTGAN (0.80) due to its higher sensitivity to rare categorical feature combinations, its MMD of 0.012 reflects superior fidelity in capturing complex tabular data distributions.

Although the Gaussian Copula achieves a lower KS score**,** TVAE obtains the lowest MMD**,** indicating superior recovery of the *joint* distribution essential for PD–LGD modelling. Unlike Quality Score, which reflects marginal alignment and penalises rare categorical combinations, MMD captures full joint distribution similarity. Because PD and LGD prediction depend on complex multivariate interactions, MMD is a more appropriate metric for this task. This supports the selection of TVAE as the augmentation method for downstream modelling, justifying the selection of TVAE for data augmentation.

D. Predictive Model Performance

The proposed TVAE–MIMO framework (trained only on synthetic-augmented training data and evaluated on real test data) exhibits substantial improvements over baseline models.

The TVAE-augmented MIMO model significantly outperforms benchmark models across both tasks. For PD, the model achieves an

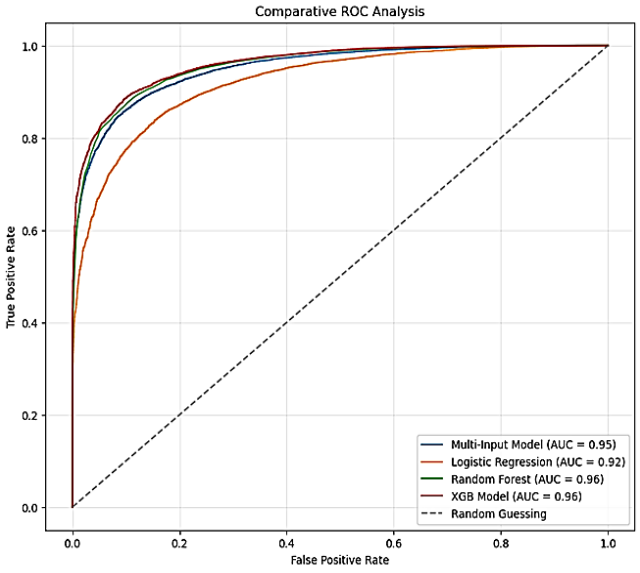


Figure 5. ROC curves for PD prediction.

AUC of 0.96, F1-Score of 0.92, and Matthews Correlation Coefficient (MCC) of 0.70, compared to Logistic Regression (AUC = 0.82, F1 = 0.75, MCC = 0.50), Random Forest (AUC = 0.87, F1 = 0.80, MCC = 0.55), and XGBoost (AUC = 0.90, F1 = 0.85, MCC = 0.60). For LGD, the model yields an R^2 of 0.76 and a MAE of 0.29, which is substantially better than Random Forest (R^2 = 0.39, MAE = 0.45) and XGBoost (R^2 = 0.50, MAE = 0.40). Figure 5 illustrates the superiority of ROC curves in predicting PD, highlighting the model's ability to discriminate between default and non-default cases effectively.

E. Feature Importance via SHAP

SHAP analysis identifies the most associative risk markers for PD and LGD predictions (Figure 6). Key markers include college year of entry (mean absolute SHAP value = 0.15), marital status (0.12), on-time graduation (0.10), unemployment rate (0.09), and loan interest rate (0.08). These findings align with established credit risk literature, confirming the socioeconomic sensitivity of default behaviour.

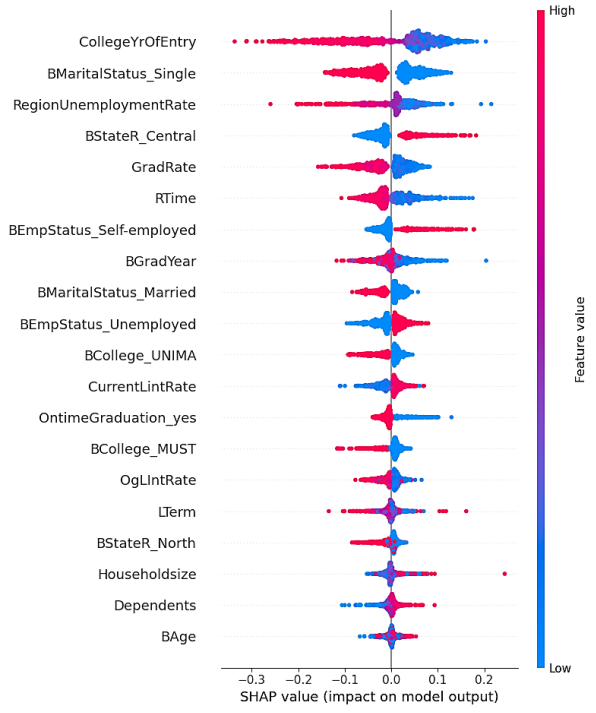


Figure 6. This beeswarm plot shows which factors most influence loan default predictions.

4. Discussions

This study examined whether generative modelling combined with multi-task neural learning can overcome the persistent challenges of data scarcity, missingness, and severe class imbalance that characterise student loan portfolios in LMICs. The empirical findings provide strong evidence that TVAE-based data augmentation, when integrated with a MIMO-MTL architecture, significantly enhances predictive accuracy for both PD and LGD. This improvement directly addresses the limitations of traditional models, which assume complete and balanced datasets and therefore underperform in settings such as those depicted by Malawi's HESLGB.

A. TVAE Effectiveness in Mitigating Data Constraints

Consistent with prior work demonstrating the robustness of VAEs for mixed-type tabular data [14, 17], TVAE achieved the lowest MMD among the evaluated generative models. It achieved the highest data fidelity with an MMD of 0.012, outperforming CTGAN (0.025) and Gaussian Copula (0.018). The lower Quality Score for TVAE is attributed to its conservative handling of rare categorical features, which prioritises robustness over exact replication of minor distributional nuances, as these features have limited impact on downstream predictive tasks. In contrast, the Gaussian Copula's higher Quality Score reflects its strength in modelling simpler, Gaussian-like distributions, but its lower Tabular Suitability limits its effectiveness for mixed data types. While the Gaussian Copula showed a lower KS statistic, its Gaussianity assumption restricted its ability to capture complex joint dependencies, leading to weaker downstream performance when applied to PD–LGD modelling. CTGAN, while effective for continuous features, struggles with categorical feature distributions, resulting in a moderate Tabular Suitability, which is a critical issue in

LMIC data where minority borrower profiles are informative but sparsely represented.

TVAE's balance of high fidelity (MMD = 0.012) and suitability for complex tabular data, despite a 2-hour training time, makes it the optimal choice for generating synthetic data in this context

B. Gains from Joint Modelling PD and LGD

It was also discovered that the TVAE-MIMO model significantly outperforms traditional methods, with an Area Under the Curve (AUC) of 0.96 for PD and an R^2 of 0.76 for LGD, compared to Logistic Regression (AUC = 0.82) and Random Forest (AUC = 0.87). Its performance reflects its ability to handle sparse, imbalanced data, validated by KS tests ($p > 0.05$) and low MMD (0.012).

The MTL framework outperformed all single-task baselines, validating the hypothesis that PD and LGD share latent socioeconomic and academic determinants. The superior joint distribution fidelity of TVAE explains the empirical gains observed in the predictive tasks: augmenting the delinquency class with TVAE samples reduced variance in model estimates and improved the separability of high-risk borrower profiles. These results confirm that high-dimensional latent learning is more effective than marginal distribution matching for educational credit risk datasets. Empirically, this manifested as a 14% AUC increase in PD prediction and a 52% improvement in LGD R^2 compared with conventional models. Practically, outperforming Logistic Regression and Random Forest indicates that the proposed model can identify high-risk borrowers earlier and with greater reliability under severe class imbalance, where conventional models systematically underestimate default risk. These improvements underscore the importance of jointly modelling PD and LGD, particularly in LMIC contexts where sparse data exacerbates estimation errors in single-task approaches.

C. Interpretation and Policy Implications for HELBs

The MIMO model's MTL approach captures the interdependencies between PD and LGD, leveraging shared socioeconomic drivers such as unemployment rates and educational outcomes, which traditional single-task models overlook [10]. SHAP analysis revealed that repayment risk is driven by temporal academic markers, such as entry year, graduation timing, marital status, regional unemployment levels, and loan cost parameters. These findings align with broader credit risk research in emerging markets, where repayment behaviour is sensitive to labour market fragility and socio-demographic constraints.

The model's enhanced ability to identify high-risk borrowers has clear policy implications, allowing HELBS to implement early interventions through targeted repayment plans before delinquency materialises. By accurately predicting at-risk borrowers, HELBs can implement targeted interventions, such as flexible repayment plans or financial literacy programs, reducing default rates and improving loan recovery. For example, SHAP suggests lowering interest rates for graduates or borrowers from stable regions, directly addressing socioeconomic disparities. Secondly, improved LGD estimation enables more accurate forecasting of expected credit losses, directly supporting the sustainability of the loan fund. The framework does not advocate risk-based exclusion. Instead, it supports socially equitable lending by enabling HELBs to anticipate vulnerabilities without reducing access for high-risk groups. Lastly, more accurate risk stratification allows loan boards to prioritise follow-ups, employer tracing, and verification processes, leading to resource optimisation. Thus, although the framework performs mathematical risk prediction, its operational impact lies in informed decision-making, not exclusionary lending. The framework

also ensures that loan programs remain viable despite data limitations.

D. Research Contribution

This study contributes to credit risk modelling in three keyways. First, it is among the earliest to combine synthetic data generation with joint PD–LGD estimation in an LMIC student loan environment, extending recent work showing that generative models can mitigate sparsity and imbalance in financial datasets [14, 17]. By integrating TVAE with a multi-task architecture, the study aligns with the evidence that default probability and loss severity are interdependent components of the credit risk process and therefore benefit from joint estimation [9]. Second, the results demonstrate that preserving the joint feature distribution measured using MMD is more predictive of downstream model performance and marginal similarity metrics such as the KS statistic. This finding is consistent with theoretical work showing that predictive accuracy depends on the full covariance structure of the data-generating process rather than univariate alignment alone. Third, the study validates the usefulness of MTL under extreme imbalance, supporting the literature that shows shared representations reduce variance when related tasks draw from limited samples [19, 20]. Generative augmentation expands the support of the minority class distribution, enabling the model to learn a smoother risk surface and improving generalisation.

The novelty of this study lies in its early application of TVAE to student loan risk modelling in LMICs, an area where generative models have been underutilised. Unlike prior work that relies on ensemble methods [10] or semi-supervised learning [26], this framework combines TVAE's high-fidelity data generation with MTL's ability to model task interdependencies, achieving a 10–14% improvement in AUC over state-of-the-art methods like XGBoost (AUC = 0.90, 2023) and attention-based neural networks (AUC = 0.91, 2024) [27]. This advancement contributes to the global knowledge base by demonstrating how generative modelling can address data scarcity in financial modelling, particularly for unsecured loans in volatile economic contexts.

E. Strengths and Limitations

The study's primary strength lies in its integration of TVAE and MTL, which addresses data scarcity and captures PD-LGD interdependencies, thereby achieving substantial predictive performance in LMIC contexts. The use of SHAP enhances transparency, enabling HELBs to understand risk drivers and tailor interventions. However, limitations exist that may impact the significance of the findings. Minor discrepancies in temporal feature correlations, such as loan amount (0.22 vs. -0.03), could impact long-term predictions in rapidly changing economic conditions, potentially underestimating default risks during economic downturns. The computational cost poses challenges for LMICs with limited infrastructure; however, cloud-based deployment mitigates this issue. The MIMO model's shared layers, while effective, introduce opacity, complicating causal inference and limiting the isolation of specific feature effects. Lastly, the model primarily suffers from generalisability. The model is trained on Malawi's HESLGB data, whose economic and institutional characteristics may not transfer directly to other LMICs. External validation through cross-country benchmarking or transfer learning would strengthen applicability. These limitations suggest that while the framework is robust, its applicability in dynamic or resource-constrained settings requires careful consideration.

Ethical considerations are paramount in deploying AI-driven credit risk models, particularly in LMICs where vulnerable populations may be disproportionately affected. TVAE's latent space abstraction minimises privacy risks by generating synthetic data that does not directly replicate individual records, reducing the potential for re-

identification. Continuous monitoring and stakeholder engagement are recommended to maintain fairness and address potential biases as economic and social conditions evolve.

F. Future Research Directions

Multiple directions for future research are evident. The primary limitation of the current framework is its static treatment of borrower behaviour. Modelling temporal dependencies in repayment patterns, particularly those influenced by economic shocks, may enhance predictive accuracy. Incorporating recurrent architectures such as LSTM-based encoders within the TVAE, or employing temporal VAEs, would enable the model to capture dynamic behavioural changes over time. Another important aspect is real-time model adaptation. Integrating current macroeconomic indicators or borrower repayment events could allow TVAE outputs to respond to evolving conditions, thereby increasing relevance in volatile LMIC contexts. Finally, combining temporal generative models with federated learning presents an opportunity for cross-country collaboration while maintaining data privacy, allowing HELBs in different jurisdictions to benefit from shared model improvements without disclosing sensitive borrower information. These enhancements would increase the scalability, robustness, and practical applicability of generative approaches to financial risk management in resource-constrained environments.

5. Conclusion

The combination of TVAE with a Multi-Task MIMO architecture offers an effective and theoretically grounded approach to student loan risk modelling in data-scarce environments. Analysis of administrative data from Malawi's HESLGB demonstrates that this framework substantially improves the prediction of default and loss severity. The results indicate that generative augmentation can compensate for the sparsity and imbalance characteristic of financial datasets in many LMICs. These findings support the emerging evidence that preserving joint feature structure is essential for downstream predictive performance and highlight the benefits of modelling interdependent credit risk components within a unified learning framework. Beyond predictive gains, the use of SHAP ensures transparency and interpretability, aligning the approach with public-sector accountability requirements and enabling loan administrators to trace the socioeconomic and academic factors shaping repayment behaviour. This potential advantage strengthens the method's practical relevance for Higher Education Loan Boards seeking to improve loan recovery and sustain equitable access to higher education.

While the study offers a robust proof-of-concept, its generalisability is currently limited to the Malawian context. Extending the framework through cross-country evaluation, transfer learning, or federated collaboration would help establish its broader applicability across diverse LMIC systems. Incorporating temporal repayment dynamics or sequence-based models can enhance robustness, particularly in volatile economic environments. Overall, the study provides a scalable, policy-relevant methodological foundation for credit risk modelling in data-scarce settings. It demonstrates that generative AI can play a meaningful role in strengthening the financial sustainability of student loan schemes while supporting fair and evidence-based decision-making. The framework thus offers a pathway for modernising risk assessment in LMIC public finance systems and advancing the application of generative models within the broader financial analytics literature.

ACKNOWLEDGEMENT

We thank the Malawi Higher Education Students Loans and Grants Board (HESLGB) for providing anonymised loan data.

CONTRIBUTIONS OF CO-AUTHORS

Reuben Moyo	[ORCID: 0009-0000-9540-5315]	Conceived the idea, conducted experiments and wrote the paper
Jimmy Mbelwa	[ORCID: 0000-0002-2263-3218]	Provided technical guidance and reviewed the paper
Juma Lungo	[ORCID: 0000-0001-6803-0757]	Provided technical guidance and reviewed the paper

REFERENCES

[1] J. K. Bii, K. Maina, and F. Sporta, 'Antecedents of Loan Recovery of Higher Education Loans in Kenya', *Ed. Cowan J. Finance Account.*, vol. **3**, no. **1**, Art. no. 1, Oct. 2023, doi: 10.55077/edithcowanjournaloffinanceandaccounting.v3i1.120.

[2] I. Chirwa, G. Masaiti, N. Mwale, S. B. Mkandawire, I. M. Mulenga, and N. K. Sichula, 'Loans and Scholarships in Africa's Higher Education Finance: A Comparative Analysis of Capitation, Policy and Recoveries in Eleven Countries', *Br. J. Educ.*, vol. **10**, no. **16**, pp. 67–88, 2022, doi: 10.37745/bje.2013/vol10n166788.

[3] Malawi Government, *Higher Education Students Loans and Grants Board Act*, vol. No **2** of 2015. Lilongwe, 2015, pp. 3–4.

[4] United Republic of Tanzania, 'The Higher Education Student's Loans Board Act, CAP. 178', no. **21**, pp. 1–30, 2008.

[5] World Bank, 'Education Finance Watch 2024', 2024. Accessed: Apr. 02, 2025. [Online]. Available: https://documents1.worldbank.org/curated/en/099102824144527868/pdf/P50097819250a00ce1812018168df2deaa3.pdf

[6] UNESCO, *Education finance watch 2023: special edition for the African Union Year of Education 2024*. GEM Report UNESCO, 2024. doi: 10.54676/FBHB3925.

[7] G. Mbewe, C. Banda, and G. Kibombwe, 'Perceptions of students on the effectiveness of higher education student loans in Malawi', *East Afr. J. Arts Soc. Sci.*, vol. **8**, pp. 1–15, May 2025, doi: 10.37284/eajass.8.2.2934.

[8] S. J. Mgaiwa and J. M. Ishengoma, 'Financing higher education in Tanzania through students' loans scheme and its impact on equitable access', *Heliyon*, vol. **9**, no. 3, p. e13943, Mar. 2023, doi: 10.1016/j.heliyon.2023.e13943.

[9] X. Dastile, T. Celik, and M. Potsane, 'Statistical and machine learning models in credit scoring: A systematic literature survey', *Appl. Soft Comput.*, vol. **91**, p. 106263, June 2020, doi: 10.1016/j.asoc.2020.106263.

[10]Y. Wang and X. S. Ni, 'A Survey of Machine Learning Methodologies for Loan Evaluation in Peer-to-Peer (P2P) Lending', in *Data Analytics for Management, Banking and Finance: Theories and Application*, F. Saâdaoui, Y. Zhao, and H. Rabbouch, Eds, Cham: Springer Nature Switzerland, 2023, pp. 1–49. doi: 10.1007/978-3-031-36570-6_1.

[11]N. Patki, R. Wedge, and K. Veeramachaneni, 'The Synthetic Data Vault', in *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, Oct. 2016, pp. 399–410. doi: 10.1109/DSAA.2016.49.

[12]M. Endres, A. Mannarapotta Venugopal, and T. S. Tran, 'Synthetic Data Generation: A Comparative Study', in *Proceedings of the 26th International Database Engineered Applications Symposium*, in IDEAS '22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 94–102. doi: 10.1145/3548785.3548793.

[13]L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, 'Modeling Tabular data using Conditional GAN', in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds, Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/254ed7d2de3b23ab10936522dd547b78-Paper.pdf

[14]S. Tazwar, M. Knobbout, E. Quesada, and M. Popa, 'Tab-VAE: A Novel VAE for Generating Synthetic Tabular Data':, in *Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods*, Rome, Italy: SCITEPRESS - Science and Technology Publications, 2024, pp. 17–26. doi: 10.5220/0012302400003654.

[15]D. Carvajal-Patiño and R. Ramos-Pollán, 'Synthetic data generation with deep generative models to enhance predictive tasks in trading strategies', *Res. Int. Bus. Finance*, vol. **62**, p. 101747, 2022, doi: https://doi.org/10.1016/j.ribaf.2022.101747.

[16]Y. Tan, H. Zhu, J. Wu, and H. Chai, 'DPTVAE: Data-driven prior-based tabular variational autoencoder for credit data synthesizing', *Expert Syst. Appl.*, vol. **241**, p. 122071, 2024, doi: https://doi.org/10.1016/j.eswa.2023.122071.

[17]F. K. Dankar, M. K. Ibrahim, and L. Ismail, 'A Multi-Dimensional Evaluation of Synthetic Data Generators', *IEEE Access*, vol. 10, pp. 11147–11158, 2022, doi: 10.1109/ACCESS.2022.3144765.

[18]S. Chatterjee, D. Corbae, M. Nakajima, and J.-V. Ríos-Rull, 'A quantitative theory of unsecured consumer credit with risk of default', *Econometrica*, vol. **75**, no. **6**, pp. 1525–1589, 2007, doi: 10.1111/j.1468-0262.2007.00806.x.

[19]R. Caruana, 'Multitask Learning', *Mach. Learn.*, vol. **28**, no. 1, pp. 41–75, July 1997, doi: 10.1023/A:1007379606734.

[20]S. Ruder, 'An Overview of Multi-Task Learning in Deep Neural Networks', June 15, 2017, *arXiv*: arXiv:1706.05098. doi: 10.48550/arXiv.1706.05098.

[21]S. Barua, D. Gavandi, P. Sangle, L. Shinde, and J. Ramteke, 'Swindle: Predicting the Probability of Loan Defaults using CatBoost Algorithm', *Int. Conf. Comput. Methodol. Commun.*, Apr. 2021, doi: 10.1109/iccmc51019.2021.9418277.

[22]K. H. Erian, P. H. Regalado, and J. M. Conrad, 'Missing data handling for machine learning models', *IJRA 2021*, pp. 123–132, 2021.

[23]X. Ma, J. Sha, D. Wang, Y. Yu, Q. Yang, and X. Niu, 'Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning', *Electron. Commer. Res. Appl.*, vol. **31**, pp. 24–39, Sept. 2018, doi: 10.1016/j.elerap.2018.08.002.

[24]M. Stenger, R. Leppich, I. Foster, S. Kounev, and A. Bauer, 'Evaluation is key: a survey on evaluation measures for synthetic time series', *J. Big Data*, vol. 11, no. **1**, p. 66, May 2024, doi: 10.1186/s40537-024-00924-7.

[25]J. R. Lanzante, 'Testing for differences between two distributions in the presence of serial correlation using the KOLMOGOROV–SMIRNOV and Kuiper's tests', *Int. J. Climatol.*, vol. **41**, no. 14, pp. 6314–6323, Nov. 2021, doi: 10.1002/joc.7196.

[26]W. Bao, N. Lianju, and K. Yue, 'Integration of unsupervised and supervised machine learning algorithms for credit risk assessment', *Expert Syst. Appl.*, vol. **128**, pp. 301–315, Aug. 2019, doi: 10.1016/j.eswa.2019.02.033.

[27]S. Kwon, J. Jang, and C. O. Kim, 'Credit scoring using multi-task Siamese neural network for improving prediction performance and stability', *Expert Syst. Appl.*, vol. **259**, p. 125327, Jan. 2025, doi: 10.1016/j.eswa.2024.125327.